

2-2019

# Why Didn't Experts Pick M4-Competition Winner?

J. Scott Armstrong
*University of Pennsylvania*, armstrong@wharton.upenn.edu

Kesten C. Green

# Why Didn't Experts Pick M4-Competition Winner?

**Abstract**

**Purpose:** Commentary on M4-Competition and findings to assess the contribution of data models—such as from machine learning methods—to improving forecast accuracy.

**Methods**: (1) Use prior knowledge on the relative accuracy of forecasts from validated forecasting methods to assess the M4 findings. (2) Use prior knowledge on forecasting principles and the scientific method to assess whether data models can be expected to improve accuracy relative to forecasts from previously validated methods under any conditions.

**Findings:** Prior knowledge from experimental research is supported by the M4 findings that simple validated methods provided forecasts that are: (1) typically more accurate than those from complex and costly methods; (2) considerably more accurate than those from data models.

**Limitations:** Conclusions were limited by incomplete hypotheses from prior knowledge such as would have permitted experimental tests of which methods, and which individual models, would be most accurate under which conditions.

**Implications:** Data models should not be used for forecasting under any conditions. Forecasters interested in situations where much relevant data are available should use knowledge models.

**Disciplines**

Business | Marketing

# Why didn't experts pick M4-Competition winner?

**J. Scott Armstrong**
jscott@upenn.edu
The Wharton School, University of Pennsylvania
and Ehrenberg-Bass Institute, University of South Australia

**Kesten C. Green**
kesten.green@unisa.edu.au
University of South Australia Business School
and Ehrenberg-Bass Institute, University of South Australia

February 10, 2019 Version 10

**Abstract**

**Purpose:** Commentary on M4-Competition and findings to assess the contribution of data models—such as from machine learning methods—to improving forecast accuracy.

**Methods**: (1) Use prior knowledge on the relative accuracy of forecasts from validated forecasting methods to assess the M4 findings. (2) Use prior knowledge on forecasting principles and the scientific method to assess whether data models can be expected to improve accuracy relative to forecasts from previously validated methods under any conditions.

**Findings:** Prior knowledge from experimental research is supported by the M4 findings that simple validated methods provided forecasts that are: (1) typically more accurate than those from complex and costly methods; (2) considerably more accurate than those from data models.

**Limitations:** Conclusions were limited by incomplete hypotheses from prior knowledge such as would have permitted experimental tests of which methods, and which individual models, would be most accurate under which conditions.

**Implications:** Data models should not be used for forecasting under any conditions. Forecasters interested in situations where much relevant data are available should use knowledge models.

"Facts that speak for themselves, talk in a very naive language"
Ragnar Frisch, Nobel Prize Lecture (1970, p. 16).

In the mid-1900s, there were two streams of thought about forecasting methods. One stream—led by econometricians—was concerned with developing "causal models" by using prior knowledge and evidence from experiments. The other was led by statisticians, who were concerned with identifying idealized "data generating processes" and with developing models from statistical relationships in data, in the expectation that the resulting models would provide accurate forecasts.

Makridakis, Spiliotis, and Assimakopoulos (2018) report that there were six machine learning (ML) models entered in the M4-Competition, and that the forecasts from those models were ranked 23, 37, 38, 48, 54, and 57. The median rank of the ML models was thus 43 out of the total of 60 models in the competition, including the competition's 10 benchmarks. The M4 authors concluded that, "The six pure ML methods that were submitted in the M4 all performed poorly, with none of them being more accurate than Comb and only one being more accurate than Naïve2" (p. 803), where Comb was the average of single, Holt, and damped exponential smoothing forecasts—the competition's "statistical benchmark"—and Naïve2 forecasts were from a random walk model with seasonal adjustment.

Makridakis, et al. (2018) make a distinction between the six ML models and the "statistical" models in the competition, but do not describe what they mean by ML methods. We understand ML models to be a kind of what we call "data model," by which we mean models that that are developed using automated processes to identify patterns in that data that are used to predict values that were unknown in the development of the model. We exclude from our definition simple extrapolation methods that have been validated for forecasting time-series data—what we understand to be at least some of Makridakis, et al.'s (2018) "statistical models"—and include methods that use automated routines to derive models that include potential "predictor" variables. Think of step-wise regression as an early data modeling method.

**1. What Can Be Concluded About Data Models From M4?**

One of the requirements of the scientific method is to test multiple reasonable hypotheses (Chamberlin 1890). In the case of the M4-Competition (Makridakis, et al., 2018), that would involve using prior knowledge to specify hypotheses on which method—i.e., *not* individual model or competition entry—would, on average, provide the most accurate forecasts under each of the conditions that would apply in the tests.

The conditions for the tests should be representative of the situations that are the subject of the study—in other words, ecologically valid. The competition's 100,000 time series are identified only by their frequency—e.g., weekly, monthly, yearly—and broad classification—e.g., macro, financial, or industrial (see Makridakis, 2018a; and Makridakis, 2018b)—and so lack the contextual information that forecasters would have available to them in practice.

The competition format is intended to attract entrants who believe that their model might be the best, and so it was necessary to acknowledge an individual model as the "winner." Drawing any scientific conclusions about the single winner would, however, require hypotheses about the relative accuracies—under specified conditions—of the 50 models that were submitted to the competition. In sum, without the relevant hypotheses, any discussion of the results for the individual models can only be speculation.

Another requirement for science is replication. Would the same model provide the most accurate forecasts for another 100,000 series? Or even for a split of the 100,000 used in the Competition? And, what if alternative error measures were used?

Regarding error measures, Makridakis et al. (2018) use two. They are the symmetrical mean absolute percentage error (sMAPE), and a combination of that measure and the mean absolute scaled error (MASE) in the form of the overall weighted average of the relative sMAPE and the relative MASE, or OWA. On the basis of the OWA, only 17 of the entries (34%) beat the competition's naïve benchmark. One of those was the entry that produced the second most accurate forecasts—on the basis of sMAPE—by using models that were developed using the principles described in Armstrong and Green (2018) (from a personal communication from Srihari Jaganathan.)

Not reported were the results on the basis of three error measures used in the M3-Competition (Makridakis and Hibon, 2000): percent better, the *median* symmetric absolute percentage error, and the median relative absolute error. Other measures that would be relevant are the cumulative relative absolute error or CumRAE (Armstrong and Collopy 1992), and the unscaled mean bounded relative absolute error or UMBRAE (Chen, Twycross, and Garibaldi, 2017.) Given that the ranking of the models is different between the two measures that *were* reported in M4—even though one of the measures (OWA) is an average that includes the first measure—it seems likely that other error measures, too, would produce different orderings of the models. For example, the model that was ranked fourth in Table 1 (Makridakis et al, 2018) provided the second most accurate forecasts on the basis of sMAPE. Different orderings would be less likely if the relative accuracy of forecasts from broad classes of *methods* were being compared, of course.

The M4-Competition provides experimental evidence that data models on average provide forecasts that are less accurate than those from simple previously validated methods. That finding is consistent with Keogh and Kasetty's (2003) conclusion from tests of data models that used diverse data sets and

performance measures. They found insufficient evidence to conclude that the methods could be useful in practice.

In the remainder of this commentary, we describe why data models *cannot* be relied upon to provide usefully accurate forecasts in typical forecasting situations in which at least some contextual and causal knowledge are available.

## 2. Do Data Models Comply with the Golden Rule of Forecasting?

The Golden Rule of forecasting is to "Be conservative by using prior knowledge about the situation and about forecasting methods." It was developed and tested by Armstrong, Green and Graefe (2015). The Golden Rule paper provides 28 evidence-based guidelines, which were tested by reviewing all papers that we could find with relevant evidence. The review identified 105 papers with 150 experimental comparisons. *All comparisons* supported the guidelines. On average, ignoring a single guideline increased forecast error by more than 40% on average. We were astonished by those findings.

One way to incorporate prior knowledge is to use the findings of experiments on which methods work best for the type of situation being forecast. In addition, in practical forecasting situations, one can use experts' domain knowledge about the expected directions of trends. In an earlier M-Competition, these two sources of knowledge were implemented by "Rule-based forecasting" (Collopy and Armstrong 1992).

Rule-based forecasting uses domain knowledge to select extrapolation models based on 28 conditions of the data in order to produce combined extrapolation forecasts. It uses 99 simple rules to weight each of the extrapolation methods. The six-year ahead *ex-ante* forecasts made by rule-based forecasting were 42% less than to those from equal-weights combinations. Other sources of prior knowledge can be used such as decomposition of time-series by level and change (Armstrong and Tessier, 2015) and by causal forces (Armstrong and Collopy, 1993.)

## 3. Do Data Models Comply with Occam's Razor?

Occam's Razor, a principle that was described by Aristotle (Charlesworth, 1956), states that one should prefer the simplest hypothesis, or model, that does the job. A review of 32 studies found 97 comparisons between simple and complex methods (Green and Armstrong, 2015.) *None* found that complexity improved forecast accuracy. On the contrary, complexity increased errors by an average of 27% in the 25 papers with quantitative comparisons.

Unsurprisingly, then, all of the validated methods for forecasting are simple. For a checklist of validated methods, see the Methods Checklist at ForecastingPrinciples.com.

## 4. Data Models Enable Advocacy, Leading to Unscientific and, Potentially, Unethical Practices

The nature of data modeling procedures is such that researchers can develop data models to provide the forecasts that they know their clients or sponsors would prefer. Doing so helps them to get grants and promotions. They can also use them to support their own preferred hypotheses.

Traditionally, researchers have justified their models by claiming that they are statistically significant, but it is not difficult to obtain statistically significant findings. And, of course, all tested relationships become statistically significant if the sample sizes are large enough.

Despite the widespread use of statistical significance testing, the technique has never been validated. On the contrary, decades of research have found that statistical significance testing harms scientific advances. For example, Koning, et al. (2005) concluded that statistical tests showed that combining forecasts did not reduce forecasting errors in the M3-Competition. Those findings were refuted in Armstrong (2007).

In the case of newer data modeling methods, statistical significance tests are not used to justify the models. Nevertheless, the complexity and opacity of the procedures, the data that are made available to the procedures, and the resulting models' makes it possible for modelers to select models that are consistent with their preferred hypotheses.

**5. Do Data Models Violate the Guidelines for Regression Analysis?**

Data models are related to regression analysis in that the models are estimated only from the data the modeling procedures are provided with. With that in mind, we rated data modeling procedures against the *Checklist for forecasting using regression analysis* at ForecastingPrinciples.com. By our ratings, typical procedures for estimating data models violate at least 15 of the 18 guidelines in the checklist. For example, they obviously violate the guidance to "select variables using prior knowledge, logic based on known relationships, and experimental studies." A summary of evidence on the limitations of regression analysis is provided in Armstrong (2012).

**6. Do data models (ML models) follow guidelines for the scientific method?**

There are eight criteria that need to be met for scientific research. They are to: (1) study an important problem, (2) use prior knowledge, (3) fully disclose hypotheses, data, and methods, (4) use an objective design, (5) use valid and reliable data, (6) use validated methods, (7) use experimental evidence, and (8) draw logical conclusions. We each independently rated "data models" using a 26-item "Compliance with Science" checklist described in Armstrong and Green (2018.) Our combined assessment was that 7.5 out of the 8 required criteria are violated by data modeling methods.

## When Data are Plentiful Use Knowledge Models, *not* Data Models

Benjamin Franklin proposed a simple method that we now refer to as "knowledge models" and, previously, as "index models." The procedure is as follows:

a. Use domain knowledge to specify:
   - all important causal variables,
   - directions of their effects, and
   - when sufficient knowledge is available, the magnitudes of their relationships.

   Variables should be scaled to relate positively to the thing that is being forecast.
   There is no limit on the number of causal variables that can be used.
   Variables may be as simple as binary, or "dummy" variables.
b. Use equal weights and standardized variables in the model unless there is strong evidence of differences in relative effect sizes. Equal weights are often more accurate than regression weights, especially when there are many variables and where prior knowledge about relative effect sizes is poor.
c. Forecast the values of the causal variables in the model.
d. Apply the model weights to the forecast causal variable values and sum to calculate a score.
e. The score is a forecast. A higher score means that the thing being forecast is likely to be better, greater, or more likely than would be the case for a lower score. If there are sufficient data to do so, estimate a single regression model that relates scores from the model with the actual values of the thing being forecast.

Evidence to date suggests that knowledge models are likely to produce forecasts that are more accurate than those from data models in situations where many causal variables are important. One knowledge model found error reductions of 10% to 43% compared to established regression models for forecasting elections in the U.S. and Australia (Graefe, Green, and Armstrong, 2019).

In conclusion, science advances not by looking for evidence to support a favorite hypothesis, but by using prior knowledge to propose hypotheses and to design experiments to test them in order to discover useful principles and methods. Our suggestion for future competitions is that when competitors submit their models and forecasts, they should use prior research to explain the principles and methods that they used, and their prior hypotheses on relative accuracy of the methods under different conditions. That would allow the competition organizers, commentators, and other researchers to compare the results by method category, taking account of prior experimental evidence, rather than resorting to *ex post* speculation on what can be learned from the results.

**References**


Armstrong, J. S. (2007). Significance Tests Harm Progress in forecasting. *International Journal of Forecasting,* 23, 321-336.

Armstrong, J. S. (2012). Illusions in regression analysis, *International Journal of Forecasting*, 28, 689-694.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8, 69–80.

Armstrong, J. S. & Collopy, F. (1993). Causal Forces: Structuring Knowledge for Time-series Extrapolation, *Journal of Forecasting, 12*, 103-115.

Armstrong, J. S., Collopy, F. & Yokum, J. T. (2005). Decomposition by Causal Forces: A Procedure for Forecasting Complex Time Series, *International Journal of Forecasting*, *21* (2005), 25-36.

Armstrong, J. S. & Green, K. C. (2018). Guidelines for Science: Evidence-based checklists. Working Paper, DOI: 10.2139/ssrn.3055874

Armstrong, J. S. & Green, K. C. (2018). Forecasting methods and principles: Evidence-based checklists, *Journal of Global Scholars in Marketing Science*, 28, 103-159.

Armstrong, J. S, Green, K. C., & Graefe, A. (2015). Golden rule of Forecasting: Be Conservative. *Journal of Business Research*, 68, 1717-1731.

Armstrong, J. S. & Tessier, T. (2015). Decomposition of time-series by level and change, *Journal of Business Research*, *68* (2015), 1755-1758.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. Reprinted in 1965 in *Science, 148,* 754-759.

Charlesworth, M. J. (1956). Aristotle's razor. *Philosophical Studies*, 6, 105–112.

Chen, C., Twycross, J., & Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, 12(3): e0174202. https://doi.org/10.1371/journal.pone.0174202

Collopy, F. & Armstrong, J. S. (1992). Rule-Based Forecasting: Development and Validation of an Expert Systems Approach to Combining Time Series Extrapolations, *Management Science*, 38,1394-1414.

Einhorn, H. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly, 36,* 367-378.

Graefe, A., Green, K. C. & Armstrong, J. S. (2019). Accuracy gains from conservative forecasting: Tests using variations of 19 econometric models to predict 154 elections in 10 countries. *PLoS ONE*, 14(1), e0209850.

Frisch, R. (1970). From Utopian Theory to Practical Applications: The Case of Econometrics. *Lecture to the Memory of Alfred Nobel, June 17, 1970*. Nobel Media AB, NobelPrize.org.

Green, K. C. & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research,* 68, 1678-1685.

Keogh, E. & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7, 349–371.

Koning, A.J., Franses, P. H., Hibon, M., & Stekler, H. O..(2005). The M3-Competition: Statistical tests of the results. *International Journal of Forecasting*, 21, 397-409.

Makridakis, S. (2018). The Dataset. *University of Nicosia*, https://www.mcompetitions.unic.ac.cy/the-dataset/
Makridakis, S. (2018). Info. *University of Nicosia*, https://www.mcompetitions.unic.ac.cy/wp-content/uploads/2018/12/M4Info.csv
Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16, 451-476.
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34, 802-808.

Words:
Text, excluding abstract = 2,132
References 467

Acknowledgements: We thank Robert Fildes, Andreas Graefe, Eamon Keogh, and an anonymous reviewer for reviewing drafts of this paper.